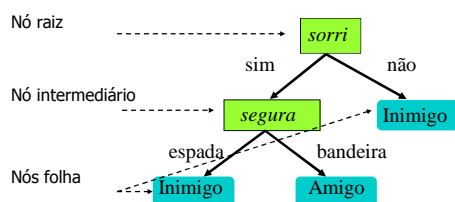
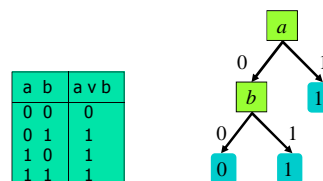
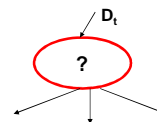
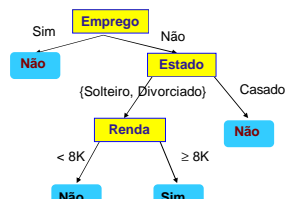
Mineração de Dados em Biologia Molecular
Árvores de Decisão
André C. P. L. F. de Carvalho
Monitor: Valéria Carvalho

Principais tópicos
Introdução
Algoritmo de Hunt
Medidas para escolha de atributos
Ponto de referência
Critério de parada
Espaço de hipóteses
25/10/2012
André de Carvalho - ICMC/USP
2

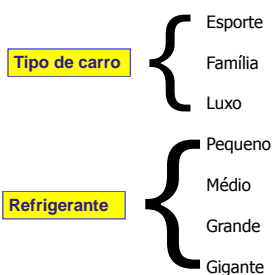
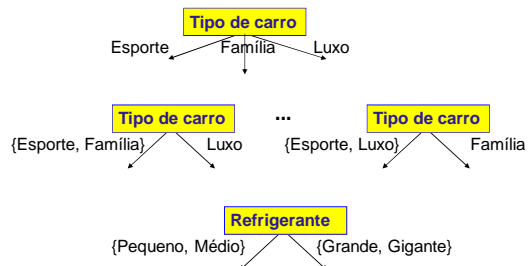
Introdução
Alguns algoritmos de AM induzem ADs a partir de um conjunto de dados
Nó raiz
Nó intermediário
Nós folha
sorri
sim
não
segura
Inimigo
espada
bandeira
Inimigo
Amigo
25/10/2012
André de Carvalho - ICMC/USP
3

Outro exemplo simples
a b | a v b
0 0 | 0
0 1 | 1
1 0 | 1
1 1 | 1
a
0
1
b
1
0
1
0
1
Nós internos e raiz: atributos preditivos
Nós externos (folhas): classes
25/10/2012
André de Carvalho - ICMC/USP
4

Exercício
Encontrar árvore de decisão para:
a AND b
a XOR b
(a AND b) OR (b AND c)
25/10/2012
André de Carvalho - ICMC/USP
5

Indução de AD
Existem vários algoritmos
Algoritmo de Hunt
Um dos primeiros
Base de vários algoritmos atuais
CART
ID3, C4.5
SLIQ, SPRINT
25/10/2012
André de Carvalho - ICMC/USP
6

Emprego	Estado	Renda	Classe
Sim	Solteiro	9500	Não
Não	Casado	8000	Não
Não	Solteiro	7000	Não
Sim	Casado	12000	Sim
Não	Divorciado	9000	Sim
Não	Casado	6000	Não
Sim	Divorciado	4000	Não
Não	Solteiro	8500	Sim
Não	Casado	7500	Não
Não	Solteiro	9000	Sim

Emprego	Estado	Renda	Classe
Sim	Solteiro	9500	Não
Não	Casado	8000	Não
Não	Solteiro	7000	Não
Sim	Casado	12000	Não
Não	Divorciado	9000	Sim
Não	Casado	6000	Não
Sim	Divorciado	4000	Não
Não	Solteiro	8500	Sim
Não	Casado	7500	Não
Não	Solteiro	9000	Sim

Indução de ADs
Geralmente usa estratégia gulosa de divisão e conquista
Divide progressivamente objetos baseado em um atributo de teste
Escolhido para otimizar algum critério
Decisões importantes
Como dividir os objetos?
Método para escolha do atributo de teste
Quando parar de dividir os objetos?
25/10/2012
André de Carvalho - ICMC/USP
13

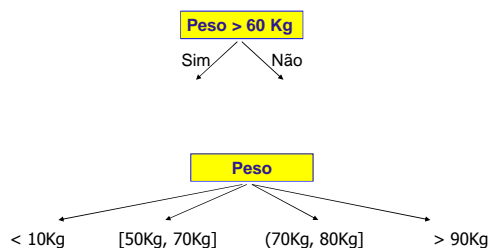
Como dividir os objetos?
Depende do tipo do atributo
Binário
Simbólico (mais que dois valores)
Nominal
Ordinal
Numérico (discreto ou contínuo)
Depende do número de divisões
2 divisões
Mais que 2 divisões
25/10/2012
André de Carvalho - ICMC/USP
14

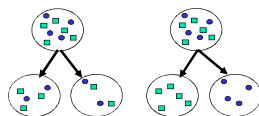
Atributos binários
Teste mais simples
Apenas dois possíveis resultados
Universidade
Pública
Privada
25/10/2012
André de Carvalho - ICMC/USP
15

Atributos simbólicos
Duas formas de condição de teste
Fazer #ramos = #possíveis valores
Agrupar parte dos valores em cada ramo
Ordinais:
Valores agrupados não devem violar relação de ordem
Nominais:
Objetos em um nó filho podem estar associados a um grupo de valores
25/10/2012
André de Carvalho - ICMC/USP
16

Atributos simbólicos
Tipo de carro
Esporte
Família
Luxo
Refrigerante
Pequeno
Médio
Grande
Gigante
25/10/2012
André de Carvalho - ICMC/USP
17

Atributos simbólicos
Tipo de carro
Esporte
Família
Luxo
Tipo de carro
{Esporte, Família}
Luxo
...
Tipo de carro
{Esporte, Luxo}
Família
Refrigerante
{Pequeno, Médio}
{Grande, Gigante}
25/10/2012
André de Carvalho - ICMC/USP
18

Atributos discretos ou contínuos

- Condição de teste pode ser expressa por:
  - Comparação simples (A < valor)
    - Escolher posição (valor) que gera melhor partição
      - Ponto de referência
  - Intervalos ($valor_{inf}$ < A < $valor_{sup}$)
    - Considerar todos os possíveis intervalos
      - Alguns intervalos adjacentes pode ser agregados

Atributos contínuos

Medidas para escolha de atributo

- Existem várias medidas para determinar o atributo que melhor divide os dados
  - Geram diferentes partições dos dados
  - Medidas de impureza
    - Distribuição de classes dos dados após divisão
      - Quanto mais balanceadas as classes em uma partição, pior

Medidas de impureza

- Baseadas no grau de impureza dos nós filhos
  - Quando maior, pior
- Exemplos de medidas de impureza
  - Entropia
  - Gini
  - Erro de classificação
  - Qui-quadrado

Medidas para escolha de atributo

$$Entropia(v) = -\sum_{i=1}^{C} p(i/v)\log_2 p(i/v)$$

$$Gini(v) = 1 - \sum_{i=1}^{C} [p(i/v)]^2$$

$$ErroClass(v) = 1 - \max_i [p(i/v)]$$

Onde:
P(i/v) = fração de dados pertencente a classe i em um nó v
C = número de classes
Considera-se que $0\log_2 0 = 0$

Exemplo

- Calcular a medida de impureza Gini para os dados abaixo:

$$Gini(v) = 1 - \sum_{i=1}^{C} [p(i/v)]^2$$

C1	0
C2	6
Gini=?	

C1	1
C2	5
Gini=?	

C1	2
C2	4
Gini=?	

C1	3
C2	3
Gini=?	

Exemplo
$Gini(v) = 1 - \sum_{i=1}^{C} [p(i/v)]^2$
P(C1) = 0/6 = 0 P(C2) = 6/6 = 1
Gini = 1 – P(C1)² – P(C2)² = 1 – 0 – 1 = 0
P(C1) = 1/6 P(C2) = 5/6
Gini = 1 – (1/6)² – (5/6)² = 0.278
P(C1) = 2/6 P(C2) = 4/6
Gini = 1 – (2/6)² – (4/6)² = 0.444
P(C1) = 3/6 P(C2) = 3/6
Gini = 1 – (3/6)² – (3/6)² = 0.500
C1 0
C2 6
Gini=0.000
C1 1
C2 5
Gini=0.278
C1 2
C2 4
Gini=0.444
C1 3
C2 3
Gini=0.500
25/10/2012
André de Carvalho - ICMC/USP
25

Exercício
Fazer os mesmos cálculos para as medidas de entropia e de erro de classificação
$Entropia(v) = -\sum_{i=1}^{C} p(i/v)\log_2 p(i/v)$
$ErroClass(v) = 1 - \max_i [p(i/v)]$
C1 0
C2 6
E=?
C1 1
C2 5
E=?
C1 2
C2 4
E=?
C1 3
C2 3
E=?
C1 0
C2 6
Class=?
C1 1
C2 5
Class=?
C1 2
C2 4
Class=?
C1 3
C2 3
Class=?
25/10/2012
André de Carvalho - ICMC/USP
26

Medida Gini média ponderada
Usada pelos algoritmos CART, SLIQ, SPRINT
Quando um nó pai é dividido em k filhos, a impureza da divisão é definida por:
$Gini_{divisão} = \sum_{t=1}^{k} \frac{N(v_f)}{N(v_p)} Gini(v_f)$
Média ponderada
Onde:
$N(v_t)$: número de objetos no filho f (v_f)
$N(v_p)$: número de objetos no nó pai (v_p)
25/10/2012
André de Carvalho - ICMC/USP
27

Medidas para selecionar divisão
Escolha da medida influencia seleção da condição de teste
Avaliação de qualidade de uma condição de teste
Comparar grau de impureza antes e após a divisão
Quanto maior a diferença, melhor a condição
Comparação pode se dar pela medida de ganho
Algoritmo ID3
25/10/2012
André de Carvalho - ICMC/USP
28

Medida de ganho
$\Delta = I(v_p) - \sum_{t=1}^{k} \frac{N(v_f)}{N(v_p)} I(v_f)$
Onde:
I(v): mede o grau de impureza do nó v
$N(v_t)$: número de objetos no filho t (v_f)
$N(v_p)$: número de objetos no nó pai (v_p)
Quando a medida de impureza é entropia, Δ mede o ganho de informação (Δ_{info})
25/10/2012
André de Carvalho - ICMC/USP
29

Divisão de atributos binários
Pai
C1 6
C2 6
Gini = 0.500
A?
Sim
Não
Nó N_1
Nó N_2
Nó 1 Nó 2
C1 4 2
C2 3 3
Gini = 0.486
B?
Sim
Não
Nó N_3
Nó N_4
Nó 3 Nó 4
C1 1 5
C2 4 2
Gini = 0.375
$Gini_{divisão}$ = (7/12)x0.49 + (5/12)x0.48 = 0.486
$Gini_{divisão}$ = tarefa de casa = 0.375
25/10/2012
André de Carvalho - ICMC/USP
30

	Tipo de Carro		
	Família	Esporte	Luxo
C1	1	2	1
C2	4	1	1
$Gini_{Div}$	???		

	Tipo de Carro	
	{Esporte, Luxo}	{Família}
C1	3	1
C2	2	4
$Gini_{Div}$	???	

	Tipo de Carro	
	{Esporte}	{Familiar, Luxo}
C1	2	2
C2	1	5
$Gini_{Div}$	???	

	Tipo de Carro		
	Família	Esporte	Luxo
C1	1	2	1
C2	4	1	1
$Gini_{Div}$	0.393		

	Tipo de Carro	
	{Esporte, Luxo}	{Família}
C1	3	1
C2	2	4
$Gini_{Div}$	0.400	

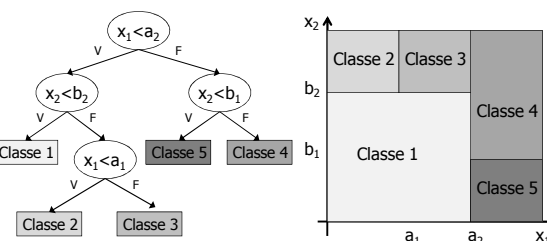
	Tipo de Carro	
	{Esporte}	{Familiar, Luxo}
C1	2	2
C2	1	5
$Gini_{Div}$	0.419	

Divisão de atributos contínuos

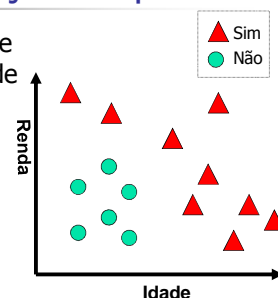
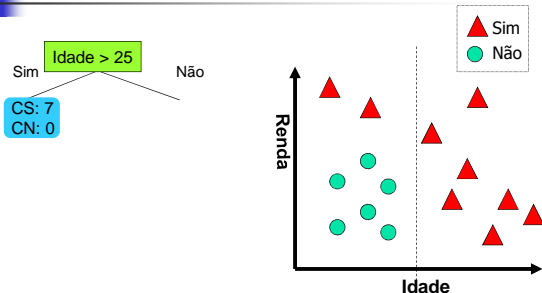
- Várias possíveis escolhas para o ponto de referência
  - # possíveis divisões = # valores distintos
- Cada ponto de referência tem uma matriz de contagens associada a ele
  - Contagens das classes em cada uma das partições
- O mesmo vale para atributos discretos

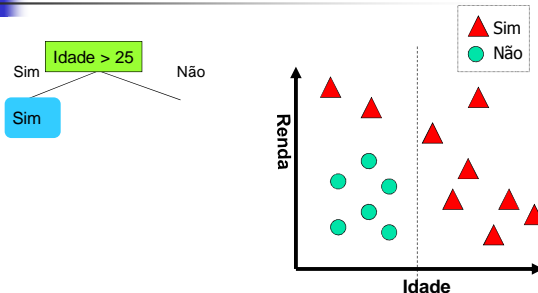
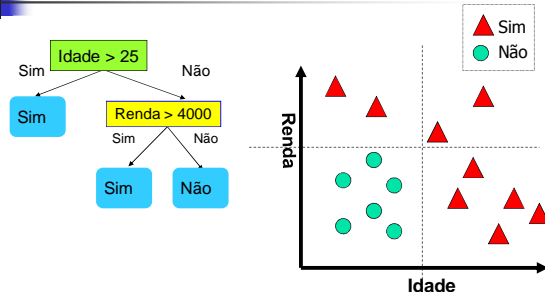
Critério de parada

- Diversas alternativas:
  - Os objetos do nó atual têm a mesma classe
  - Os objetos do nó atual têm valores iguais para os atributos de entrada, mas classes diferentes
  - O número de objetos no nó é menor que um dado valor
  - Todos os atributos já foram incluídos no caminho

Espaço de hipóteses
Cada percurso da raiz a um nó folha representa uma regra de classificação
Cada folha
Está associada a uma classe
Corresponde a uma região do espaço de soluções
Hiper-retângulo
Interseção de hiper-retângulos é um conjunto vazio
União é o espaço total
25/10/2012
André de Carvalho - ICMC/USP
37

Espaço de hipóteses
x1<a2
V
F
x2<b2
x2<b1
Classe 1
x1<a1
Classe 5
Classe 4
Classe 2
Classe 3
x2
b2
b1
Classe 2
Classe 3
Classe 4
Classe 1
Classe 5
a1
a2
x1
25/10/2012
André de Carvalho - ICMC/USP
38

Exemplo
Sejam os dados abaixo referentes a solicitações de crédito bancário
Construir uma árvore de decisão que classifica aplicação para cartão de crédito
Idade Renda Classe
20 2000 Sim
30 5200 Não
60 5000 Sim
40 6000 Não
...
25/10/2012
André de Carvalho - ICMC/USP
39

Busca no espaço de hipóteses
Construir uma AD que classifica solicitante de cartão de crédito
Aprova (Sim)
Não aprova (Não)
Sim
Não
Renda
Idade
25/10/2012
André de Carvalho - ICMC/USP
40

Busca no espaço de hipóteses
Idade > 25
Sim
Não
CS: 7
CN: 0
Sim
Não
Renda
Idade
25/10/2012
André de Carvalho - ICMC/USP
41

Busca no espaço de hipóteses
Idade > 25
Sim
Não
Sim
Sim
Não
Renda
Idade
25/10/2012
André de Carvalho - ICMC/USP
42

Nome	Febre	Enjôo	Manchas	Dores	Diagnóstico
João	sim	sim	pequenas	sim	doente
Pedro	não	não	grandes	não	saudável
Maria	sim	sim	pequenas	não	saudável
José	sim	não	grandes	sim	doente
Ana	sim	não	pequenas	sim	saudável
Leila	não	não	grandes	sim	doente

Perguntas
25/10/2012
André de Carvalho - ICMC/USP
49